ESTUDI ALCOHOL

ÍNDEX

1. Introducció

Segons la OMS, l'alcohol es tracte del factor causal de més de 200 malalties i cada any es produeixen més de 3 milions de defuncions degut a aquest consum (5.3% del total). Tot i que el seu consum de forma ocasional i en poca quantitat pot no ocasionar problemes greus en la salut, pot portar conseqüències a curt termini en cas de passar-se com ara: Accidents de tràfic, caigudes, cremadures, cometre o patir actes violents, etc. En cas d'abusar de la substància fins arribar al punt de caure en l'alcoholisme, els efectes poden arribar a ser molt més greus com ara l'empitjorament de les malalties mentals existents o inclús l'aparició de noves, produir malalties hepàtiques, trastorns digestius, produir danys en el cervell, nervis i produir complicacions neurològiques, problemes oculars, etc. I per tant, cal conscienciar més a la població dels efectes en la salut que poden portar l'ús inapropiat de l'alcohol.

2. Objectiu i hipòtesi de l'estudi

L'objectiu de l'estudi és identificar les diferències sociodemogràfiques, físiques i bioquímiques d'analítiques de sang de consumidors i no consumidors d'alcohol. També es pretén definir un model senzill per tal de predir si a partir d'aquestes característiques l'individu és consumidor d'alcohol o no.

Com que l'alcohol afecta principalment al fetge, podem esperar que nivells alts de la gamma GTP (un enzim que es troba en tot el cos, però que quan el fetge es troba en molt mal estat es filtra en la sang) poden ser un bon indicador del mal estat d'aquest òrgan i per tant sospitar del elevat consum. El pes i la mesura de la cintura també podrien ser bons indicadors, ja que els consumidors molt habituals d'alcohol acostumen a tenir més greix en el cos i per tant, també es podria veure afectat el nivell total de colesterol.

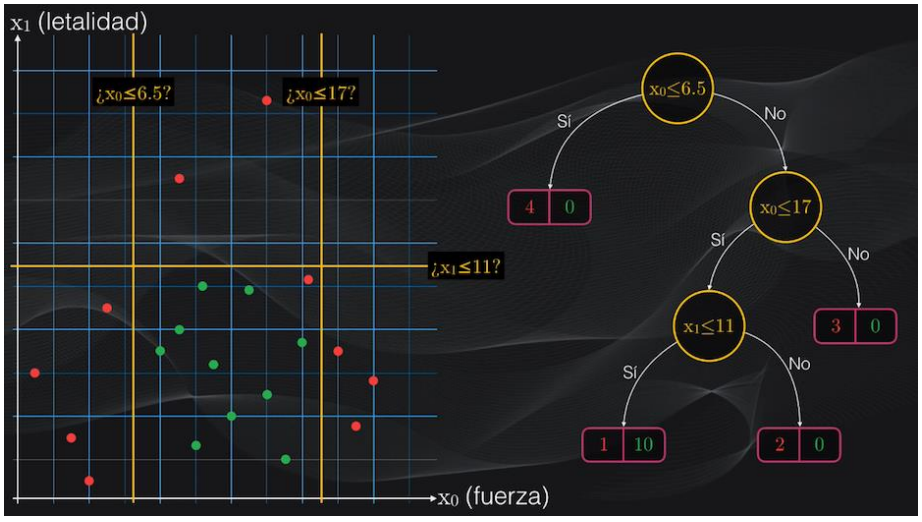
3. Material i mètodes

Per identificar les característiques que fan diferenciar els dos grups, es calcularan les mitjanes amb les seves respectives desviacions i freqüències amb els seus respectius percentatges, depenent de la variable a descriure. A més es duran a terme proves t de Student i prova d'independència chi quadrat, per identificar diferència i associacions estadístiques.

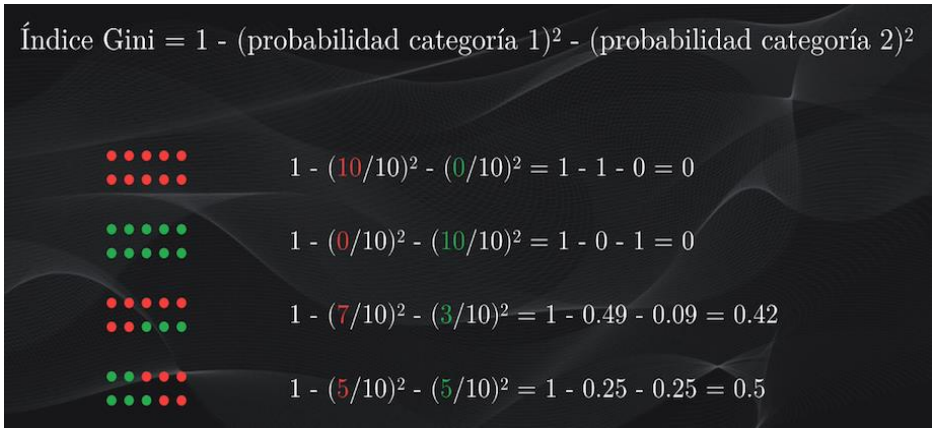
Finalment, pels models estadístics de classificació que s'utilitzaran són: l'arbre de classificació, Bagged Tree i Random Forest. Es farà ús del software estadístic R per a dur a terme l'anàlisi.

3.1 Arbre de Classificació

La idea dels arbres de classificació és molt simple: iterativament es generen particions binàries, buscant que en cada nova partició de la regió es generi un subgrup de dades el més homogeni possible (Imatge 1 esquerra). Aquestes iteracions se sol representar en forma d'arbre on el punt de partida es denomina "arrel" i conté la primera condició on es representa gràficament com dues fletxes indicant si compleix o no amb la condició. Després estan els nodes interns que segueixen generant les particions. Per mesurar aquesta homogeneïtat, s'utilitza **l'índex de Gini** que mesura el grau d'impuresa del node. Un valor de Gini igual a 0 indica nodes purs (dades que pertanyen a una sola categoria), en cas contrari indiquen nodes amb impureses. És a dir, amb dades pertanyents a més d'una categoria (veure exemple en la Imatge 2).



Imatge 1: Exemple il·lustratiu d'arbre de classificació amb dues variables (x1: Letalitat i x2: Força).



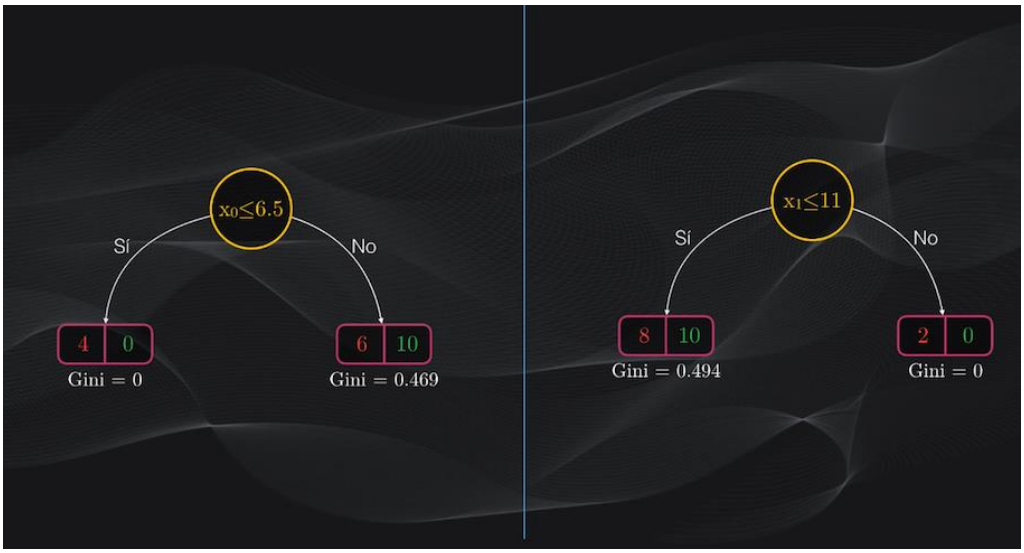
Imatge 2: Exemple il·lustratiu índex de Gini aplicat als arbres de classificació.

L'arbre assigna una puntuació utilitzant **la funció de cost** a cada node pare (Xo i X1) utilitzant la ponderació dels valors de Gini dels nodes fills (Sí – No) per a escollir la millor partició. Per exemple, calculem la funció de cost per a les dues possibles particions que es mostren en la Imatge 3:

$$Cost_{x0} = 0 \cdot \frac{4+0}{20} + 0.469 \cdot \frac{6+10}{20} = 0.375$$

$$Cost_{x1} = 0.494 \cdot \frac{8+10}{20} + 0 \cdot \frac{2+0}{20} = 0.446$$

D'entre aquests dos valors de cost, s'escull la que té un valor menor ja que indicaria un nivell d'impureses menor i per tant, una millor capacitat de classificació. Aquest procediment s'aplica de forma iterativa fins a tenir tot l'arbre construït aconseguint les agrupacions més homogenis possibles.



Imatge 3: Exemple il·lustratiu funció de cost. Node pare: X0 - X1. Node fills: Sí – No.

Per tal de reduir el sobre ajustament del nostre arbre, el que es fa és realitzar el que s'anomena el **procés de poda**. Aquest procés consisteix en eliminar alguns nodes de l'arbre entrenat, minimitzant l'error de predicció. Un dels mètodes més utilitzats és el **podatje de complexitat de costos**, que ve definit per la següent funció:

$$R_{Cp}(T) = R(T) + Cp|T|$$

On $R_{Cp}(T)$ és la suma de l'error del arbre, R(T), més el paràmetre de penalització Cp (quantificació de la compensació) multiplicat per la mida de l'arbre, |T|. L'objectiu és trobar l'arbre que minimitzi $R_{Cp}(T)$, a partir del valor de cp. Com més petit sigui el paràmetre de penalització, major serà la mida de l'arbre.

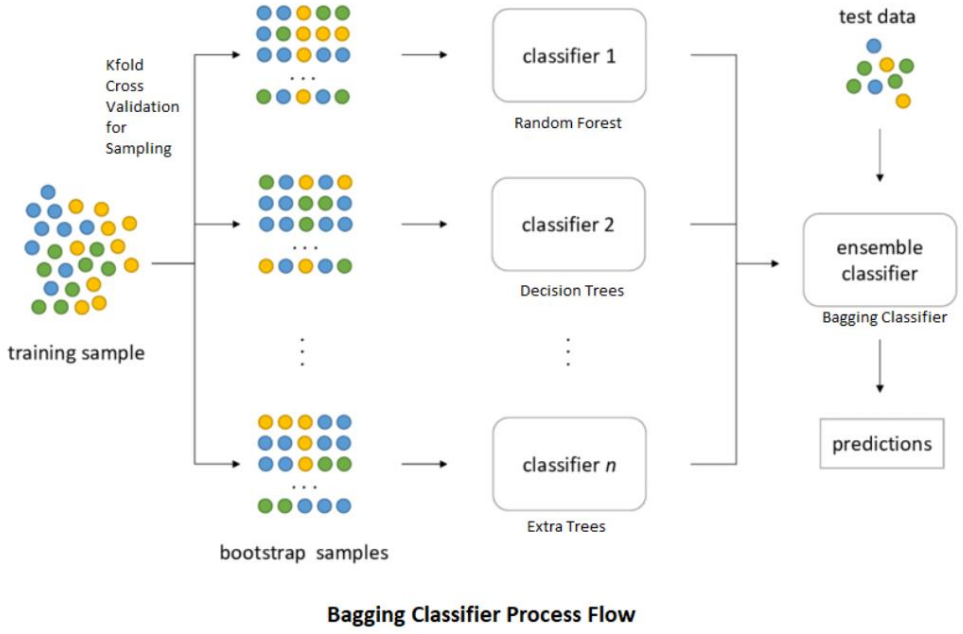
Un dels avantatges d'aquests models és que es consideren les interaccions i a més són interpretables.

.

3.2 Bagged Tree

Aquest mètode s'utilitza quan volem reduir la variància d'un arbre de decisió. Aquest mètode construeix B arbres de decisió independents utilitzant conjunts d'entrenament obtinguts mitjançant remostreig i després es combinen els resultats. En el cas dels models de classificació, s'escull la categoria més votada (la moda) i per als models de regressió la mitjana de les prediccions. Aquests arbres creixen profundament i no es poden (Imatge 4).

Per a un conjunt d'entrenament de mida *n,* cada arbre es composa de ~ $\sim (1 - e^{-1})$n observacions únics i les observacions que no s'han utilitzat per al remostreig, s'utilitzen per avaluar la precisió del model i la capacitat global del mètode s'obté promitjant la capacitat predictiva de cada arbre. Un desavantatge que té el mètode és que si cada arbre té una capacitat predictiva deficient, el rendiment promig continuarà sent deficient. Un altre limitació que comporta, és que no existeix un únic arbre amb un conjunt de regles a interpretar i com a conseqüència, no queda clar quines variables són més importants.



Imatge 4: Explicació il·lustrativa de la classificació Bagging Tree.

3.3 Random Forest

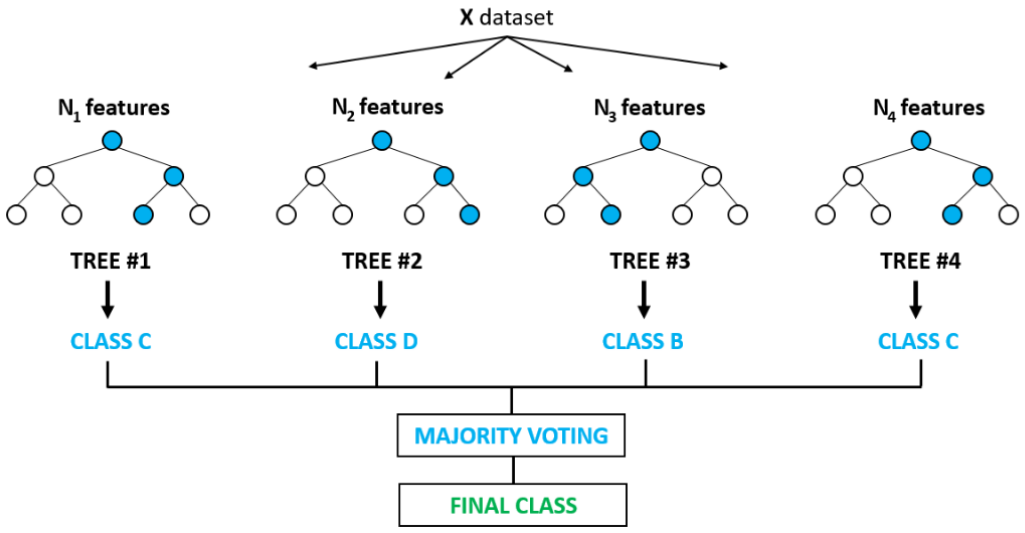
Els random forests (boscos aleatoris) també són un conjunt d'arbres de decisió que milloren els Bagged Tree mitjançant la creació d'un bosc d'arbres no correlacionades que millora la capacitat predictiva d'un sol arbre. Cada arbre del model es construeix de la següent forma (Imatge 5):

➔ Seleccionem una mostra d'N casos de forma aleatòria i amb reemplaçament. Aquesta mostra s'utilitzarà per a construir l'arbre í-essima.

➔ Si denominem M al número total de variables predictores, seleccionarem de forma aleatòria un número m<M de variables i crearem un arbre complet utilitzant només aquestes variables. Aquest valor m es mantindrà constant duran la generació de tot el bosc.

➔ Tot el bosc creix en la seva màxima extensió sense procés de poda i la predicció es fa a partir de les prediccions generades pels B arbres (majoria de vots per al model de classificació i promig de les prediccions per al model de regressió).

Sovint, aquests models s'utilitzen per reduir el nombre de variables necessaris per tal de reduir la càrrega de la recol·lecció de dades i millorar l'eficiència.



Imatge 5: Exemple il·lustratiu del mètode Random Forest.

3.4 Base de dades

La base de dades s'ha extret de la web data.go.kr, pertanyent a la National Health Insurance Service de Corea del Sud (NHIK): https://www.data.go.kr/data/15007122/fileData.do) . Es té un total de 991.346 individus amb dades bioquímiques de sang per a cada un d'ells, juntament amb algunes característiques físiques. Les variables són:

- **Sex:** Sexe del pacient (Male/Female).
- **Age**: Edat del pacient [anys].
- **Height**: Altura del pacient [cm].
- **Weight**: Pes del pacient [kg].
- **Waistline**: Mida de la cintura [cm].
- **Sight left**: Mesura l'agudesa de l'ull esquerra [diòptries].
- **Sight right**: Mesura l'agudesa de l'ull dret [diòptries].
- **Hear_left:** Agudesa oïda esquerra (1: Normal, 2: Anormal).
- **Hear_right:** Agudesa oïda dreta (1: Normal, 2: Anormal).
- **SBP:** Pressió arterial sistòlica [mmHg].
- **DBP:** Pressió arterial diastòlic [mmHg].
- **BLDS:** Nivell de sucre en sang en dejú [mg/dL].
- **tot_chole:** Nivell total de colesterol [mg/dL].
- **HDL_chol:** Nivell de colesterol LDH [mg/dL].
- **LDL_chole:** Nivell de colesterol LDL [mg/dL].
- **Triglyceride:** Nivell de triglicèrids en sang [mg/dL].
- **Hemoglobin:** Nivell d'hemoglobina en sang [g/dL].
- **urine_protein:** Proteïna en l'orina [1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)].
- **serum_creatinine:** Nivell de creatinina en sang [mg/dL].
- **SGOT_AST:** Nivells d'Aspartat en sang [IU/L].
- **SGOT_ALT:** Nivells d'Alanina en sang [IU/L].
- **gamma_GTP:** Nivells de γ-glutamyl transpeptidase [IU/L].
- **SMK_stat_type_cd:** Si el pacient no és fumador (1), ex-fumador (2) o si encara fuma (3).
- **DRK_YN:** Si el pacient és consumidor d'alcohol o no (Y: Sí, N: No).

4. Validació de les dades

La base de dades no conté cap valor faltant, per tant, fixem-nos tan sols en el rang de valors que prenen les variables (Taula 1). En cas de què es tractin de valors que sí es poden obtenir en la realitat, es mantindran. En cas contrari, s'eliminaran els casos.

		Pre validació	Post validació
age	**min**	20.00	20.00
	max	85.00	85.00
height	**min**	130.00	130.00
	max	190.00	190.00
weight	**min**	25.00	25.00
	max	140.00	140.00
waistline	**min**	8.00	8.00
	max	**999.00**	**149.10**
sight_left	**min**	0.10	0.10
	max	**9.90**	**2.50**
sight_right	**min**	0.10	0.10
	max	**9.90**	**2.50**
SBP	**min**	**67.00**	**70.00**
	max	273.00	273.00
DBP	**min**	**32.00**	**33.00**
	max	185.00	185.00
BLDS	**min**	25.00	25.00
	max	852.00	852.00
tot_chole	**min**	30.00	30.00
	max	2344.00	2344.00
HDL_chole	**min**	1.00	1.00
	max	8110.00	8110.00
LDL_chole	**min**	1	1.00
	max	5119	**5119.00**
triglyceride	**min**	1.00	1.00
	max	9490.00	9490.00
hemoglobin	**min**	1.00	1.00
	max	25.00	25.00
serum_creatinine	**min**	0.10	0.10
	max	98.00	98.00
SGOT_AST	**min**	1.00	1.00
	max	9999.00	9999.00
SGOT_ALT	**min**	1.00	1.00
	max	7210.00	7210.00
gamma_GTP	**min**	1.00	1.00
	max	999.00	999.00

Taula 1: Valor mínim i màxim per a cada covariable on en la primera columna tenim els valors originals i en la segona columna els valors després de la filtració.

En la base de dades, s'han detectat 26 casos repetits i que s'han acabat eliminant. En la taula 1, observem que el pes mínim és de 25kg. Hi ha 9 casos de dones de 85 anys amb altures de 130 i 140 cm, i per tant aquest pes pot ser possible. Per a l'edat, altura, pes, mida de la cintura, pressió arterial (SBP i DBP), no es detecten possibles valors erronis. Per a l'agudesa visual, s'ha eliminat els valors de 9.9 per ser un valor molt irreal i per a la mida de la cintura eliminem el valor de 999 cm, ja que en tots els casos es tracten de dones d'entre 20 i 40 anys amb altures d'entre 150 i 170cm i pesos d'entre 45 i 90kg i per tant no té gaire sentit. Per a la resta, cal d'algun professional per a poder valorar si es tracten de valors que es poden donar o no en la realitat. També s'ha marcat en groc els valors que han patit canvis degut a la filtració.

5. Processament de les dades

Primer eliminem aquelles variables amb molt poca variabilitat, és a dir, eliminem aquelles files que contenen TRUE a les columnes (o en només en una d'elles) zeroVar (variància zero) i nzv (variància propera a zero). En el nostre cas, tenim que l'agudesa de l'oïda (tant el dret com l'esquerra) i el nivell de proteïna en la orina es poden eliminar (Taula 2).

	freqRatio	percentUnique	zeroVar	nzv
sex	1.132.243	0.0002017459	FALSE	FALSE
age	1.007.347	0.0014122214	FALSE	FALSE
height	1.020.092	0.0013113484	FALSE	FALSE
weight	1.004.780	0.0024209509	FALSE	FALSE
waistline	1.092.102	0.0743433675	FALSE	FALSE
sight_left	1.068.757	0.0024209509	FALSE	FALSE
sight_right	1.091.992	0.0024209509	FALSE	FALSE
hear_left	30.751.521	0.0002017459	FALSE	TRUE
hear_right	31.812.988	0.0002017459	FALSE	TRUE
SBP	1.091.325	0.0172492752	FALSE	FALSE
DBP	1.102.570	0.0128108652	FALSE	FALSE
BLDS	1.000.454	0.0502347314	FALSE	FALSE
tot_chole	1.018.946	0.0478137805	FALSE	FALSE
HDL_chole	1.044.715	0.0224946689	FALSE	FALSE
LDL_chole	1.002.459	0.0435771164	FALSE	FALSE
triglyceride	1.003.534	0.1671464857	FALSE	FALSE
hemoglobin	1.008.179	0.0191658614	FALSE	FALSE
urine_protein	30.313.614	0.0006052377	FALSE	TRUE
serum_creatinine	1.079.036	0.0184597507	FALSE	FALSE
SGOT_AST	1.016.524	0.0572958382	FALSE	FALSE
SGOT_ALT	1.012.294	0.0599185350	FALSE	FALSE
gamma_GTP	1.005.521	0.0948205773	FALSE	FALSE
SMK_stat_type_cd	2.815.750	0.0003026189	FALSE	FALSE
DRK_YN	1.000.747	0.0002017459	FALSE	FALSE

Taula 2: Variabilitat de les covariables (zeroVar i nzv),
on queda marcat en groc aquelles en què aquesta variabilitat és molt baixa.

Un altre aspecte que s'ha de tenir en compte és la correlació entre les covariables, on es considerarà com a llindar el valor de 0.80. És a dir, passat aquest valor s'eliminarà. Per tant, eliminem el colesterol total (tot_chole), ja que té una correlació de 0.88 amb els nivells d'LDL (Taula 3 marcat en groc).

	age	height	weight	waistline	sight_left	sight_right	SBP
age	1	0	-0,19533671	0,1271703	-0,172096429	-0,167683633	0,26553
height	-0,39850118	1	1	0,26394524	0,139141348	0,13852891	0,03503
weight	-0,19533671	0,66882349	1	1	0,088900973	0,088707494	0,25077
waistline	0,1271703	0,26394524	0,63717313	1	0	0,006157671	0,27232
sight_left	-0,17209643	0,13914135	0,08890097	0,0045114	1	0	-0,03562
sight_right	-0,16768363	0,13852891	0,08870749	0,00615767	0,307984696	1	0
SBP	0,26552985	0,03503012	0,25077043	0,27232266	-0,035617127	-0,033993568	1
DBP	0,10884678	0,10877957	0,2778907	0,24088997	-0,001208634	-0,000568033	0,74113
BLDS	0,19579583	0,02126597	0,13858669	0,17551941	-0,034817337	-0,036892643	0,18314
tot_chole	0,01144622	-0,02323965	0,0632379	0,06320104	0,004370658	0,003436716	0,06856
HDL_chole	-0,10462419	-0,14859911	-0,28768815	-0,253988	-0,004224426	-0,006258695	-0,11177
LDL_chole	0,02949652	-0,01545036	0,06785867	0,0634163	0,003013444	0,002154707	0,03362
triglyceride	0,04354936	0,13761061	0,28377397	0,24943638	0,010599135	0,012265125	0,186
hemoglobin	-0,17308078	0,53189841	0,49949058	0,29172963	0,085896009	0,086847198	0,16653
serum_creatinine	0,02281868	0,17125646	0,15388453	0,09842143	0,020308541	0,021446628	0,0626
SGOT_AST	0,05940769	0,03920301	0,09965174	0,09685849	-0,005426815	-0,004600016	0,08148
SGOT_ALT	-0,02050554	0,14484237	0,27644021	0,21226693	0,018815082	0,018920122	0,11761
gamma_GTP	0,01739127	0,16233979	0,22188103	0,18698752	0,013566471	0,01649069	0,16143

	DBP	BLDS	tot_chole	HDL_chole	LDL_chole	triglyceride	hemoglobin
age	0,10884678	0,19579583	0,01144622	-0,104624189	0,02949652	0,04354936	-0,17308078
height	0,10877957	0,02126597	-0,023239648	-0,148599111	-0,0154504	0,13761061	0,53189841
weight	0,2778907	0,13858669	0,063237902	-0,287688146	0,06785867	0,28377397	0,49949058
waistline	0,24088996	0,17551941	0,063201039	-0,25398802	0,0634163	0,24943638	0,29172963
sight_left	-0,0012086	-0,03481734	0,004370658	-0,004224426	0,00301344	0,01059914	0,08589601
sight_right	-0,000568	-0,03689264	0,003436716	-0,006258695	0,00215471	0,01226513	0,0868472
SBP	0,74113088	0,183141	0,068557195	-0,111771566	0,03361923	0,18600294	0,16653022
DBP	1	0,13626636	0,111914935	-0,093838477	0,06698422	0,19865096	0,24198015
BLDS	0,13626636	1	0,012712886	-0,113151731	-0,0297363	0,20474714	0,10171195
tot_chole	0,11191493	0,01271289	1	0,1638685	0,87736727	0,27068333	0,12127179
HDL_chole	-0,0938385	-0,11315173	0,1638685	1	0,02208907	-0,268366	-0,18186469
LDL_chole	0,06698422	-0,02973625	0,877367272	0,022089065	1	0,02957086	0,10164596
triglyceride	0,19865096	0,20474714	0,27068333	-0,268365966	0,02957086	1	0,24164518
hemoglobin	0,24198015	0,10171195	0,12127179	-0,18186469	0,10164596	0,24164518	1
serum_creatinine	0,05705317	0,04354151	-0,005975861	-0,083896589	0,00240145	0,06019287	0,13928531
SGOT_AST	0,07855473	0,0679132	0,032409529	-0,034109927	0,00199606	0,10625955	0,09889224
SGOT_ALT	0,13131992	0,11218198	0,075957289	-0,117597032	0,04703388	0,18761397	0,22931806
gamma_GTP	0,17560997	0,16897067	0,094537112	-0,055708725	-0,0085436	0,29902655	0,22621798

	serum_creatinine	SGOT_AST	SGOT_ALT	gamma_GTP
age	0,022818681	0,05940769	-0,0205055	0,017391274
height	0,171256458	0,039203015	0,14484237	0,162339793
weight	0,153884531	0,09965174	0,27644021	0,221881026
waistline	0,098421431	0,096858492	0,21226693	0,186987521
sight_left	0,020308541	-0,005426815	0,01881508	0,013566471
sight_right	0,021446628	-0,004600016	0,01892012	0,01649069
SBP	0,06260454	0,081478013	0,11761299	0,161433643
DBP	0,057053173	0,078554726	0,13131992	0,175609972
BLDS	0,043541513	0,067913199	0,11218198	0,168970667
tot_chole	-0,005975861	0,032409529	0,07595729	0,094537112
HDL_chole	-0,083896589	-0,034109927	-0,117597	-0,055708725
LDL_chole	0,002401448	0,001996057	0,04703388	-0,008543624
triglyceride	0,060192869	0,106259553	0,18761397	0,299026552
hemoglobin	0,13928531	0,098892236	0,22931806	0,226217983
serum_creatinine	1	0,025943954	0,04784695	0,056775441
SGOT_AST	0,025943954	1	0,64172205	0,328496541
SGOT_ALT	0,047846947	0,641722053	1	0,374446471
gamma_GTP	0,056775441	0,328496541	0,37444647	1

Taula 3: Matriu de correlacions entre les covariables, on queda marcat en groc la parella altament correlacionada.

Com que tenim dades balancejades, és a dir, tenim (casi) la mateixa proporció d'individus per a cada grup, no cal utilitzar cap tècnica de balanceig (Taula 4). Finalment, al entrenar els models predictius, s'estandarditzaran les dades degut a les diferents escales de mesura dels predictors.

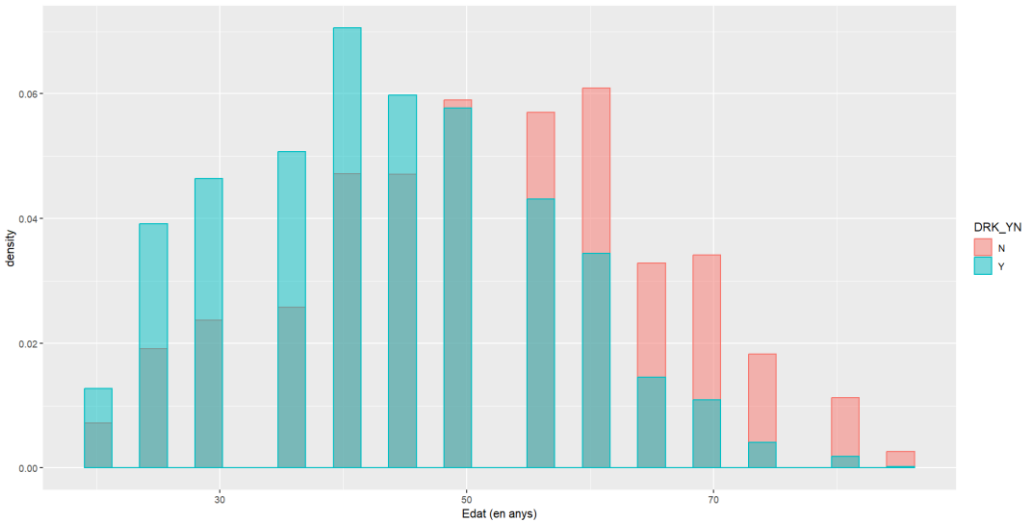
Consumidor (DRK_YN)	
N (No)	**S (Sí)**
495858 (50.02%)	495488 (49.98%)

Taula 4: Nombre d'individus de la base de dades que sí són consumidors d'alcohol (columna S) i d'individus no consumidors d'alcohol (columna N).

6. Anàlisi descriptiu

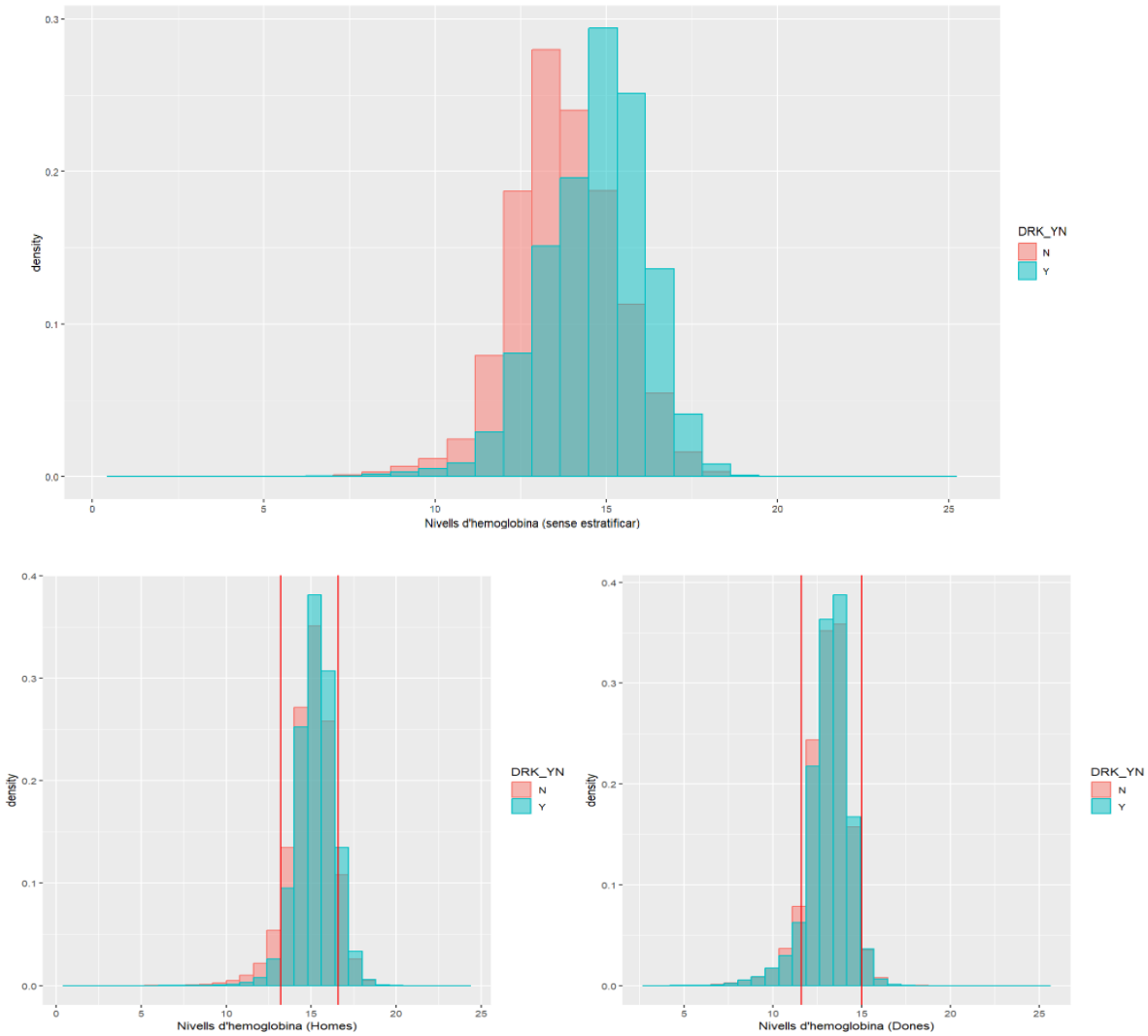
Es descriuran només aquelles variables en els que hi trobem una discrepància entre els consumidors d'alcohol i els que no ho són. En la primera gràfica (Gràfica 1), observem que els joves són més propensos a ser consumidors, on el pic màxim es troba en els 40 anys d'edat, i la mitjana es troba en els 43.5 anys (σx =12.77). On a partir dels 40 anys aquest consum comença a reduir-se. Pels no consumidors la mitjana és de 51.65 (σx =14.38) (Taula 5). Les mitjanes d'edat per a cada grup són significativament diferents (p-valor<0.001).

Els nivells d'hemoglobina (Gràfica 2) són més altes en el grup consumidor, on les mitjanes són de 14.70 (1.49) (grup consumidor) i 13.76 (1.53) g/dL (grup no consumidor), sent aquests valors significativament diferents (p-valor<0.001). Ara, si analitzem quin grup té més nivells d'hemoglobina anòmals hauríem de fer els anàlisis estratificant per sexes, ja que aquests valors depenen del sexe, i que s'han marcat en línies vermelles l'interval on es consideren valors normals. En els homes observem que la majoria dels que tenen valors anòmalament inferiors són el grup no consumidor ($\bar{x}$ ($\sigma_{\bar{x}}$) = 15.04 (1.29)) i els que tenen valors anòmalament superiors són el grup consumidor ($\bar{x}$ ($\sigma_{\bar{x}}$) = 15.32 (1.10)) i en dones no s'observen gaires diferències entre els dos grups (p-valor estratificació <0.001).



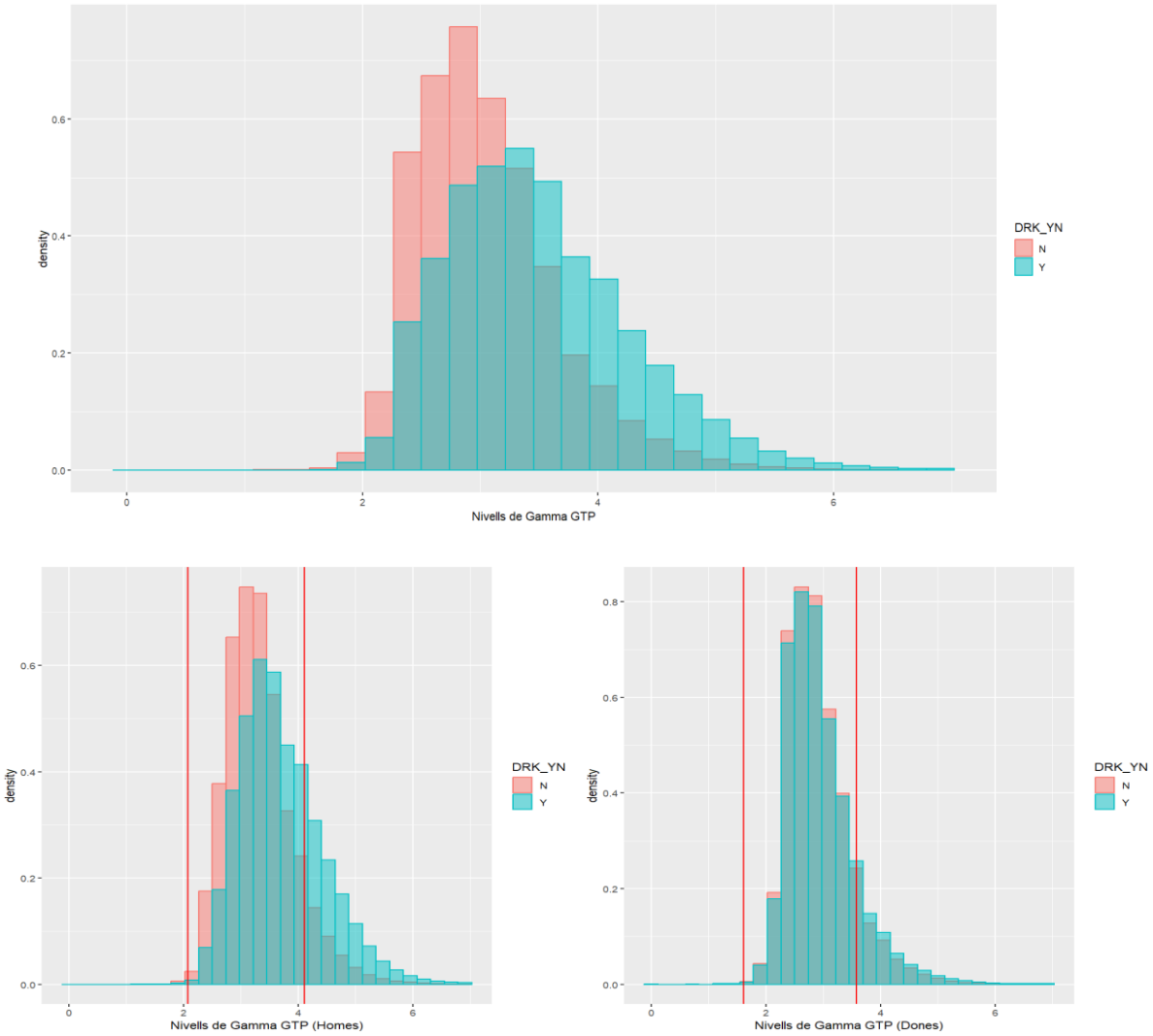
Gràfica 1: Distribució de l'edat estratificat pels dos grups d'estudi.

On en vermell es representa al grup no consumidor i en blau al grup consumidor.



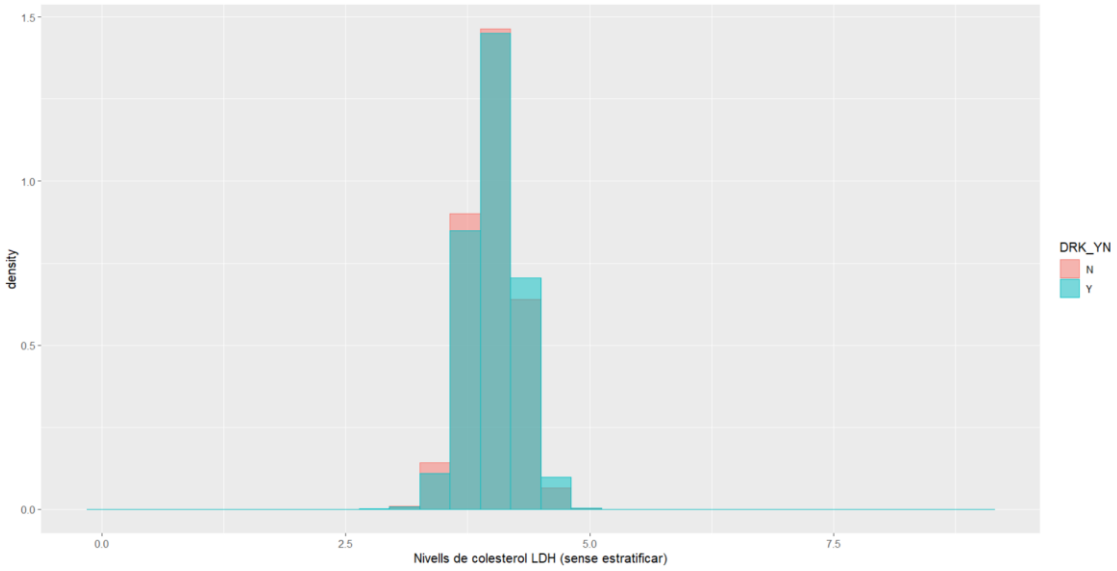
Gràfica 2: El primer gràfic es mostra els nivells d'hemoglobina en sang (en escala logarítmica) sense estratificar per sexes. A sota a l'esquerra tenim representat els nivells per als homes i a la dreta per a les dones, on queda marcat l'interval, en vermell, on cauen els valors normals.

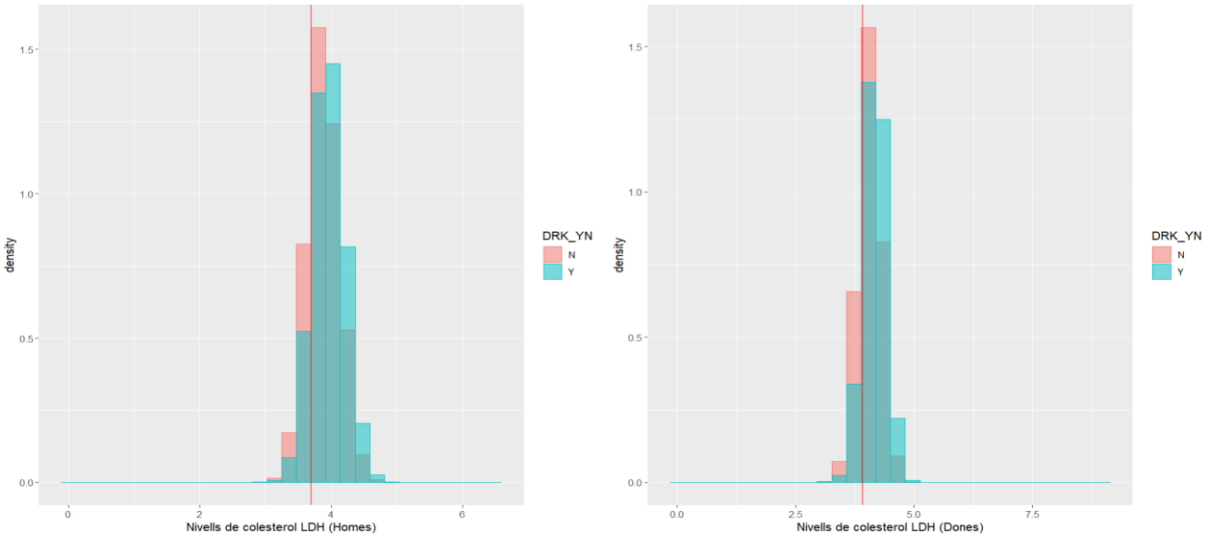
Per als nivells de gamma GTP (Gràfica 3), sense estratificar, observem que el grup consumidor tenen valors molt més elevats que els que no ho són amb mitjanes i desviacions de 26.77 (30.37) i 47.43 (62.68), respectivament. On aquests diferències resulten ser significativament diferents (p-valor<0.001). Si estratifiquem per sexes, observem que els homes consumidors tenen més valors anòmals (cua dreta, $\bar{x}$ ($\sigma_{\bar{x}}$) = 56.41 (68.43)) que els que no ho són ($\bar{x}$ ($\sigma_{\bar{x}}$) = 34.65 (37.17)). En el cas de les dones, no notem diferències en les distribucions, tot i que les mitjanes sí són significativament diferents (p-valor<0.001), 22.60 (25.07) pels no consumidors i 24.96 (36.51) pels consumidors.



Gràfica 3: En el primer gràfic es mostren els nivells de gamma GTP.

A sota, es representen els nivells estratificat per sexe, on a l'esquerra hi tenim el grup d'homes i a la dreta al grup dones.

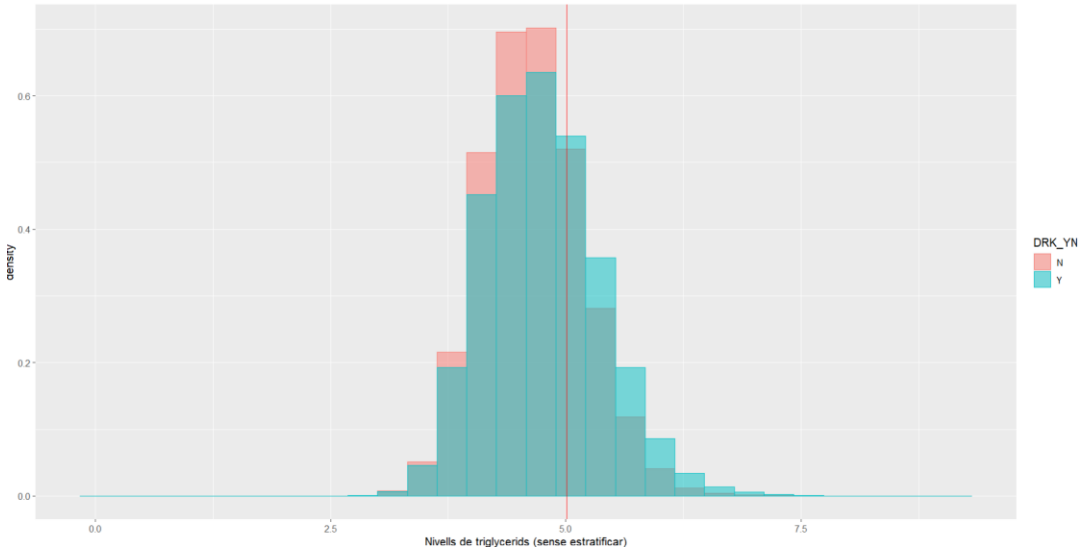




Gràfica 4: Nivells de colesterol LDH, a la part superior es mostra les distribucions sense estratificar per sexes. En la part inferior, a la dreta es mostra la distribució de les dones i a l'esquerra la distribució dels homes.

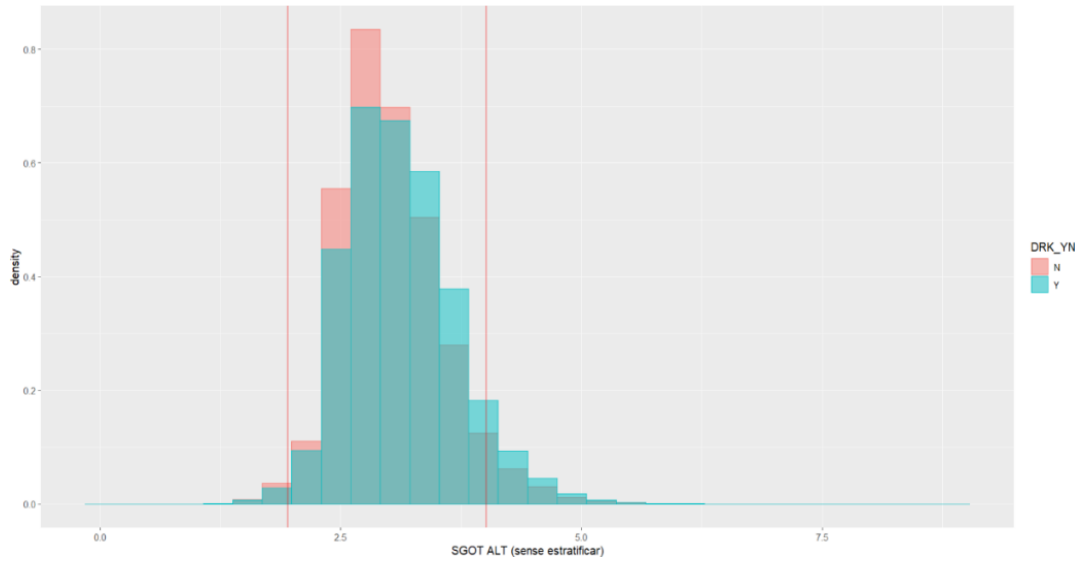
Per als nivells de colesterol LDH (Gràfic 4), en general no s'observen diferències en les distribucions, però sí en les mitjanes, on la mitjana del grup consumidor és de 57.68 (15.50) i per als no consumidors aquesta mitjana és redueix a 56.23 (18.82), sent aquestes diferències significatives (p-valor<0.001). Però si estratifiquem per sexes i ens fixem en la part de la distribució a l'esquerra de la línia vermella (que indiquen valors anòmals), observem que els no consumidors tenen valors més inferiors que els consumidors tant pels homes (consumidors $\bar{x}$ ($\sigma_{\bar{x}}$) = 54.21 (13.74); no consumidors $\bar{x}$ ($\sigma_{\bar{x}}$) = 49.85 (12.73)) com per a les dones (consumidors $\bar{x}$ ($\sigma_{\bar{x}}$) = 66.37 (16.21); no consumidors $\bar{x}$ ($\sigma_{\bar{x}}$) = 59.61 (20.57)), on aquestes diferències resulten ser significatives (p-valor<0.001)

Pels nivells de triglicèrids (Gràfica 6), observem que pel grup consumidor tenim una distribució lleugerament desplaçada cap a la dreta ($\bar{x}$ ($\sigma_{\bar{x}}$) = 121.41 (81.66)) en comparació al grup no consumidor ($\bar{x}$ ($\sigma_{\bar{x}}$) = 142.77 (118.33)) en el què es presenten diferències significatives entre les mitjanes (p-valor<0.001). Els valors normals corresponen a la part dreta de la línia vermella, on observem que hi predomina la categoria consumidor.



Gràfica 6: Nivells de triglicèrids sense estratificar per sexes.

Pels nivells d'alanina (Gràfica 7), també observem un desplaçament en la distribució dels consumidors ($\bar{x}$ ($\sigma_{\bar{x}}$) = 27.41 (29.02)) respecte als no consumidors (($\bar{x}$ ($\sigma_{\bar{x}}$) = 24.11 (23.24), sent aquests diferències estadísticament significatives. On observem més valors anòmals en els consumidors (cua dreta).



Gràfica 7: Nivells d' Alanina sense estratificar per sexes.

	categoria	N	Y	p-valor
n		495844	495476	
Edat		51.65 (14.38)	43.58 (12.77)	**<0.001**
Sexe (%)	Dona	323751 (65.3)	141170 (28.5)	**<0.001**
	Home	172093 (34.7)	354306 (71.5)	
Fumador (%)	1 (No fumador)	389000 (78.5)	213431 (43.1)	**<0.001**
	2 (ex fumador)	54471 (11.0)	120473 (24.3)	

	3 (Fumador)	52373 (10.6)	161572 (32.6)	
Hemoglobina				
	Tots	**13.76 (1.53)**	**14.70 (1.49)**	**<0.001**
	Dona	13.08 (1.18)	13.15 (1.16)	**<0.001**
	Home	15.04 (1.29)	15.32 (1.10)	**<0.001**
Gamma GTP	**Tots**	**26.77 (30.37)**	**47.43 (62.68)**	**<0.001**
	Dona	22.60 (25.07)	24.96 (36.51)	**<0.001**
	Home	34.65 (37.17)	56.41 (68.43)	**<0.001**
Colesterol LDH	**Tots**	**56.23 (18.82)**	**57.68 (15.50)**	**<0.001**
	Dona	59.61 (20.57)	66.37 (16.21)	**<0.001**
	Home	49.85 (12.73)	54.21 (13.74)	**<0.001**
Triglicèrids		121.41 (81.66)	142.77 (118.33)	**<0.001**
Alanina		24.11 (23.24)	27.41 (29.02)	**<0.001**
Visió esquerra		0.93 (0.67)	1.03 (0.53)	**<0.001**
Visió dret		0.93 (0.67)	1.02 (0.52)	**<0.001**

Taula 5: Mitjana de les variables Sexe, Fumador, nivells d'hemoglobina, nivells de gamma GTP i agudesa visual (dreta i esquerra). Entre parèntesi, s'indica la desviació estàndard.

Si mirem per sexes, els homes representen el 71.5% dels consumidors, mentre que les dones, només un 28.5%. I dins del grup no consumidor, predominen les dones, concretament un 65.3% (associació: p-valor<0.001). En el cas del estatus de fumador, entre els no consumidors predominen els individus que no fumen, concretament un 78.5%, i en el grup consumidor també però amb molt menys freqüència (43.1% i significatiu amb p-valor<0.001). Tenim més o menys el doble d' ex-fumadors en el grup consumidor (24.3%) que en el grup no consumidor (11.0%) i en quant als fumadors, predominen en el grup de consumidors amb un 32.6% (p-valor<0.001).

Tot i que les distribucions no se solapen, i per tant no podem assegurar diferències significatives entre elles, podem sospitar que aquestes variables podrien ser importants a l'hora de discriminar els grups en els diferents models predictius, a més de la variable sexe que sembla també tenir certa influència sobre aquestes.

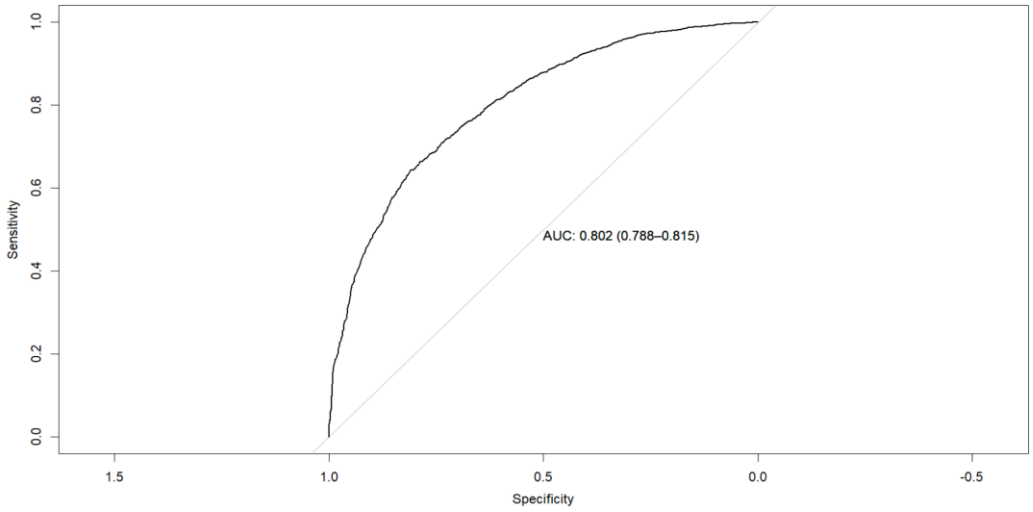
7. Entrenament dels models

Al tenir una base de dades de dimensió molt gran (991.320 individus), entrenar els models serà massa costos a nivell computacional, i per tant, es reduirà agafant tan sols el 2% dels casos. És a dir, s'agafaran en total 19.712 individus i es partiran en dos conjunts train (80%) i test (20%).

Si comencem ajustant el model Random Forest i utilitzant 400 arbres, obtenim la matriu de confusió de sota (Taula 6) on **l'accuracy és de 71.9%**, la sensibilitat és de 73.2% i l'especificitat de 70.6%. El valor **AUC és de 80.2%** i per tant tenim un bon model classificador (Gràfica 8).

Confusion Matrix and Statistics			
		Reference	
		N	**Y**
Prediction	**N**	1390	530
	Y	578	1444

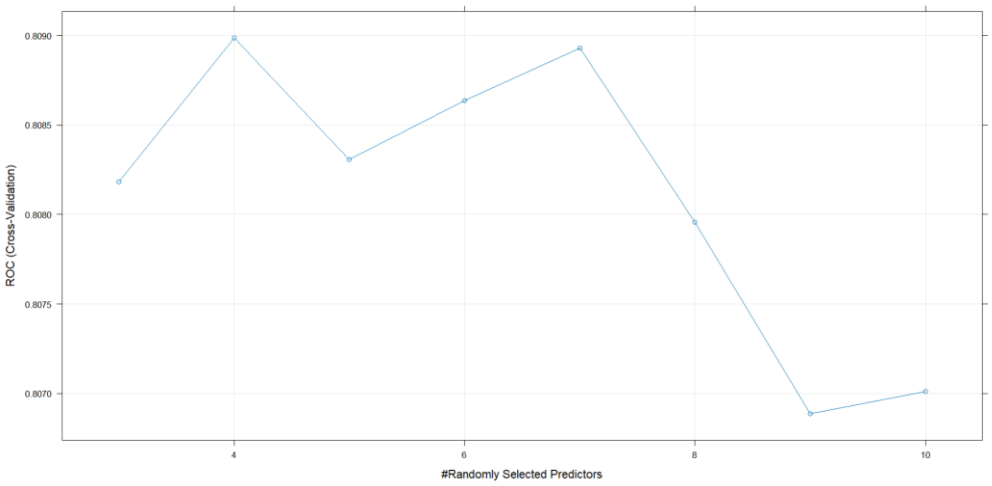
Taula 6: Matriu de confusió model Random Forest amb tots els predictos.



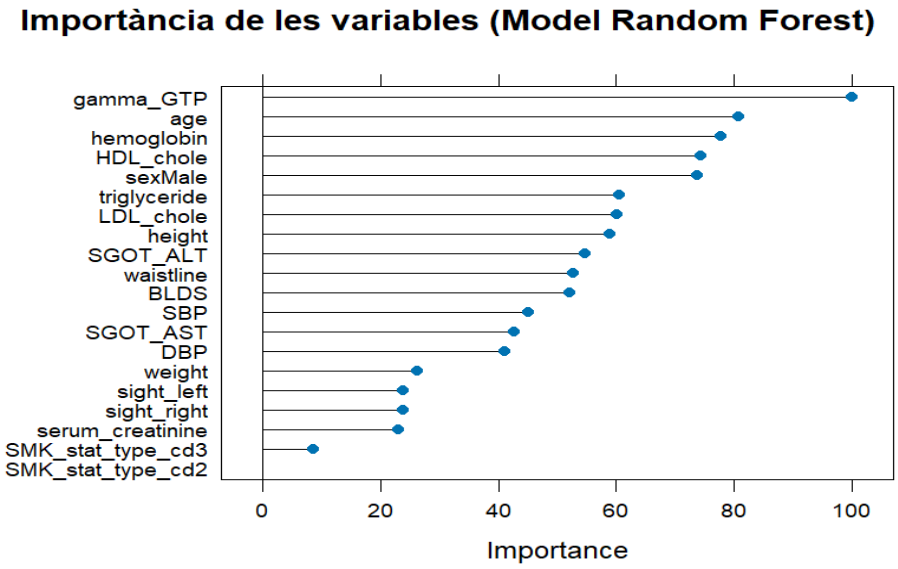
Gràfica 8: Curva ROC on s'hi representa el valor AUC del model Random Forest complet.

Si mirem la relació entre el nombre de predictors seleccionats i els valors ROC (Gràfica 9), veiem que el nombre de predictors que ens dona el model més eficient inclouen els

següents: Nivell de gamma GTP, edat, nivells d'hemoglobina i nivells de colesterol LDH (Gràfica 10).



Gràfica 9: Relació entre el nombre de predictors seleccionats pel Random Forest i els valors ROC.

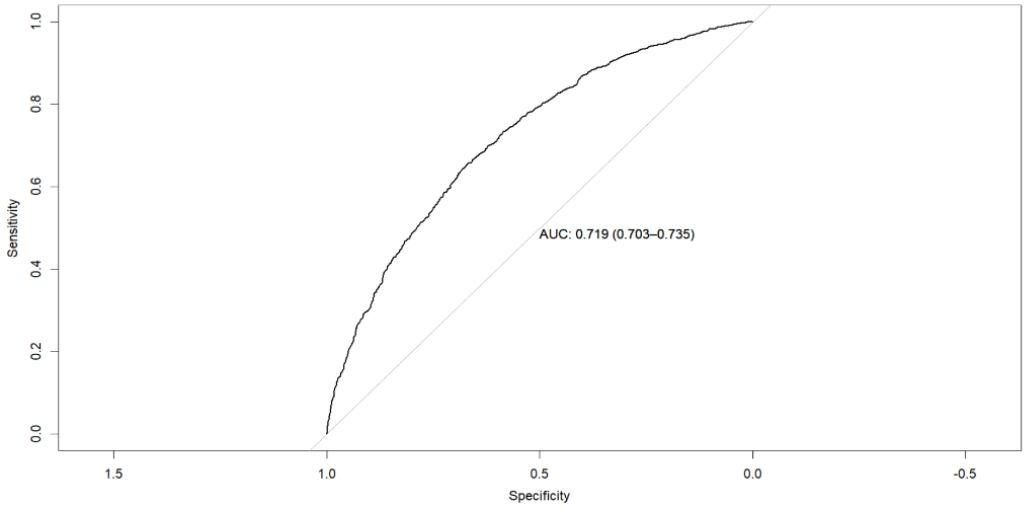


Gràfica 10: Covariables més importants segons el model Random Forest.

Ara, si reajustem el model però amb només les variables esmentades obtenim un **accuracy de 66.2%,** una sensibilitat de 65.8% i una especificitat de 66.7%. En el cas de **l'AUC**, aquest disminueix a **71.9%** (Gràfica 11). Però tot i així continua sent un bon classificador.

Confusion Matrix and Statistics			
		Reference	
		N	Y
Prediction	N	1313	676
	Y	655	1298

Taula 7: Matriu de confusió model Random Forest amb només les variables seleccionades.

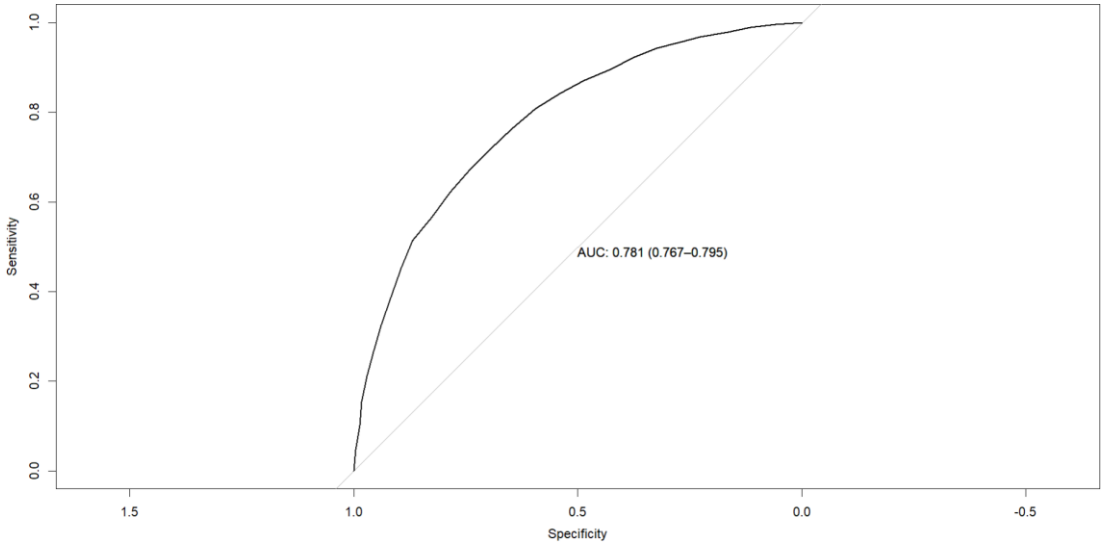


Gràfica 11: Curva ROC on s'hi representa el valor AUC del model Random Forest amb només la selecció dels 4 predictors.

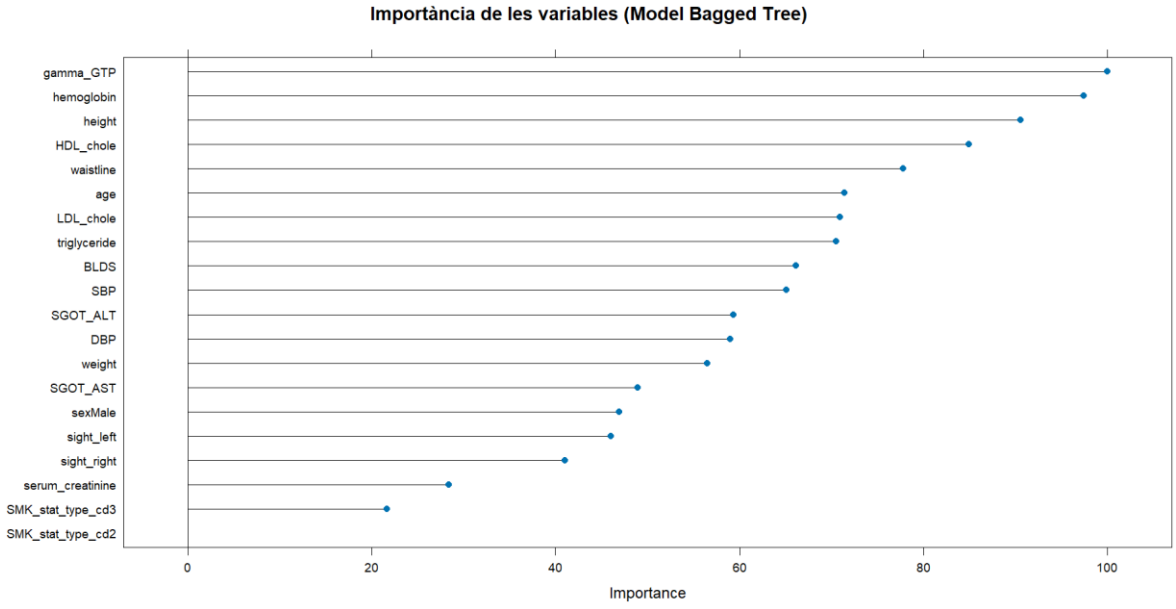
En el cas d'ajustar el model Bagged Tree, amb totes les variables obtenim un **accuracy de 70.6%**, una sensibilitat de 71.5% i una especificitat de 69.7%. En el cas de **l'AUC**, aquest pren un valor de **78.1%** (Gràfica 12), que també ens informa de que es tracte d'un bon classificador. I les variables que considera ser les més importants són: Gamma GTP, nivells d'hemoglobina, altura i nivells de colesterol LDH (Gràfica 13). Si tornem a ajustar el mode però amb les variables seleccionades pel model Random Forest, **l'accuracy** es reduiex a **64.7%**, la sensitivitat a 63.8% i l'especificitat a 65.6%. Finalment, el valor **d'AUC és de 70.5%**, indicant un model amb bones capacitats per a classificar els grups (Gràfica 14).

Confusion Matrix and Statistics			
		Reference	
		N	Y
Prediction	N	1372	562
	Y	596	1412

Taula 8: Matriu de confusió model Bagged Tree amb tots els predictos.



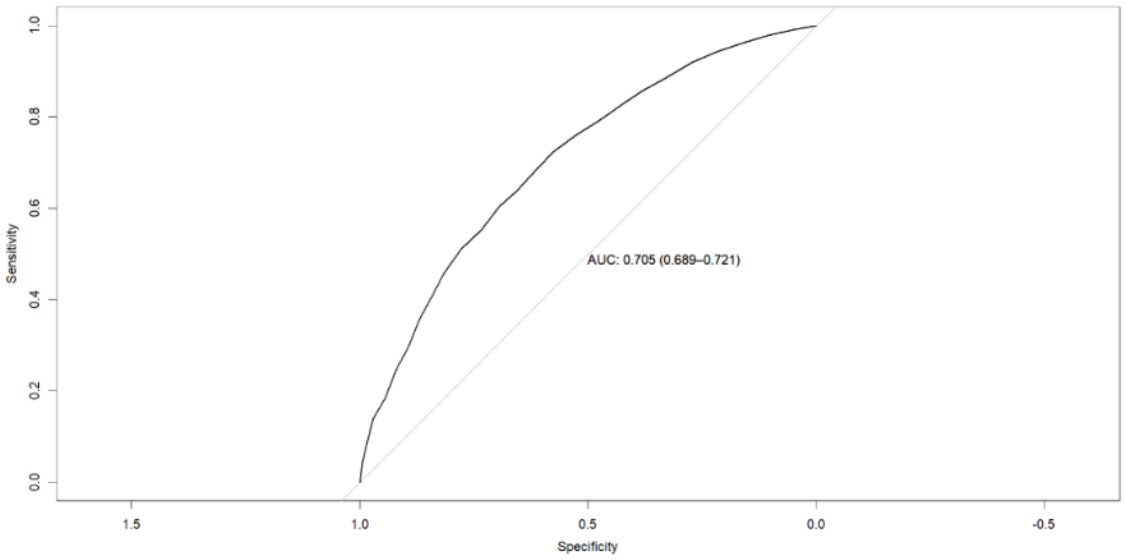
Gràfica 12: Curva ROC on s'hi representa el valor AUC del model Bagged Tree complet.



Gràfica 13: Importància de les covariables segons el model Bagged Tree.

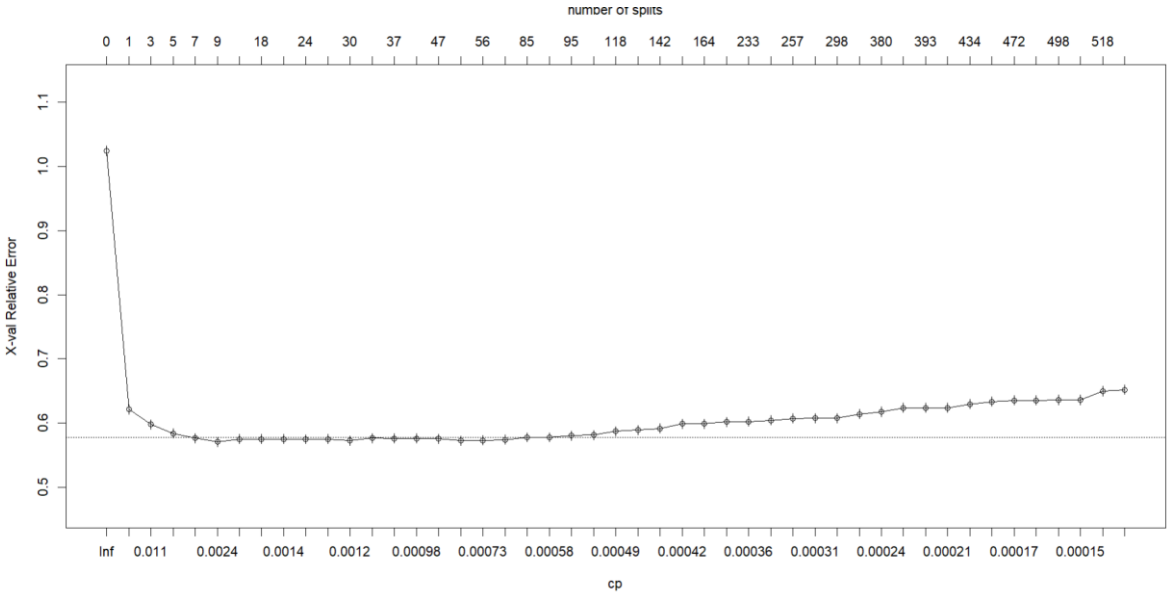
Confusion Matrix and Statistics			
		Reference	
		N	Y
Prediction	N	1290	715
	Y	678	1259

Taula 9: Matriu de confusió model Bagged Tree amb els predictors seleccionats pel model Random Forest.



Gràfica 14: Curva ROC on s'hi representa el valor AUC del model Bagged Tree amb els predictors seleccionats pel Random Forest.

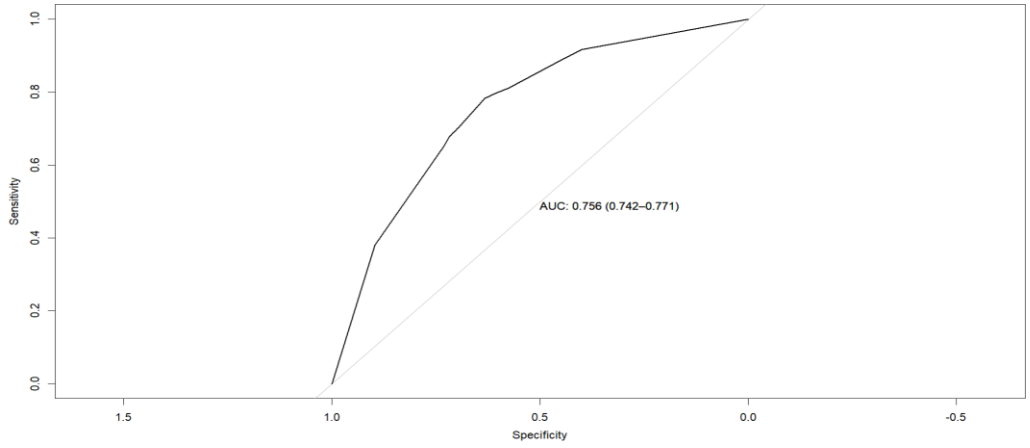
Ara, ajustem l'arbre de classificació amb tots els predictors. Si mirem la relació entre l'error de l'arbre i el nombre de particions que s'han de realitzar per a obtenir el mínim error i per evitar el sobre ajustament del model, hauríem de podar l'arbre en 9 particions. És a dir, amb un valor de cp igual a 0.0014 aproximadament, obtenim un error de 0.57 (Gràfica 15). Per a aquest model, **l'accuracy** és de **70.8%**, la sensibilitat de 78.3% i l'especificitat de 63.3%. Finalment, obtenim un **AUC** de **75.6%** (Gràfica 16).



Gràfica 15: Relació entre l'error i el nombre de particions (i Cp) per a l'arbre de classificació amb tots els predictors.

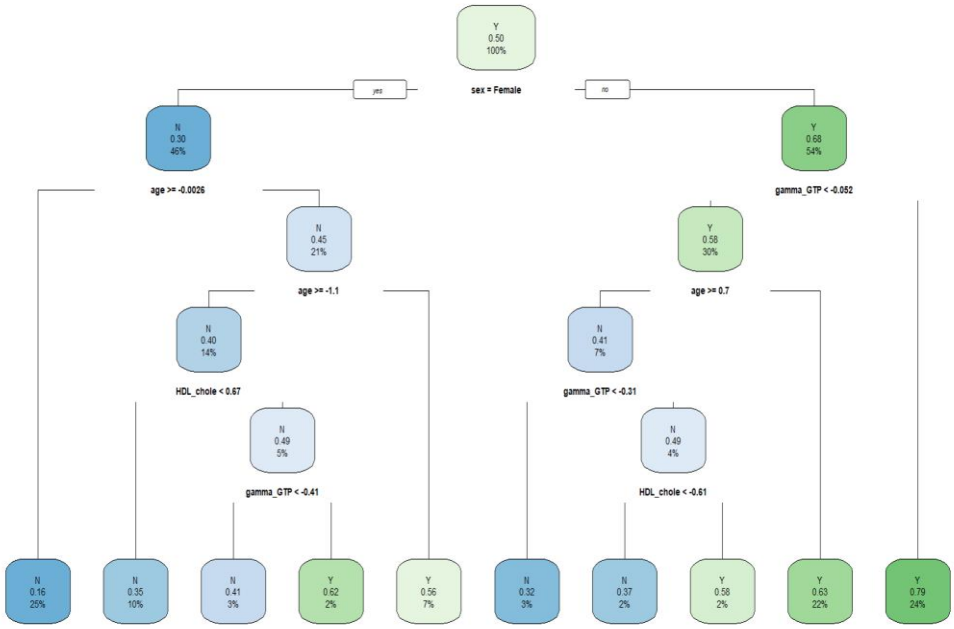
Confusion Matrix and Statistics			
		Reference	
		N	Y
Prediction	N	1245	429
	Y	723	1545

Taula 10: Matriu de confusió de l'arbre de classificació amb tots els predictors.



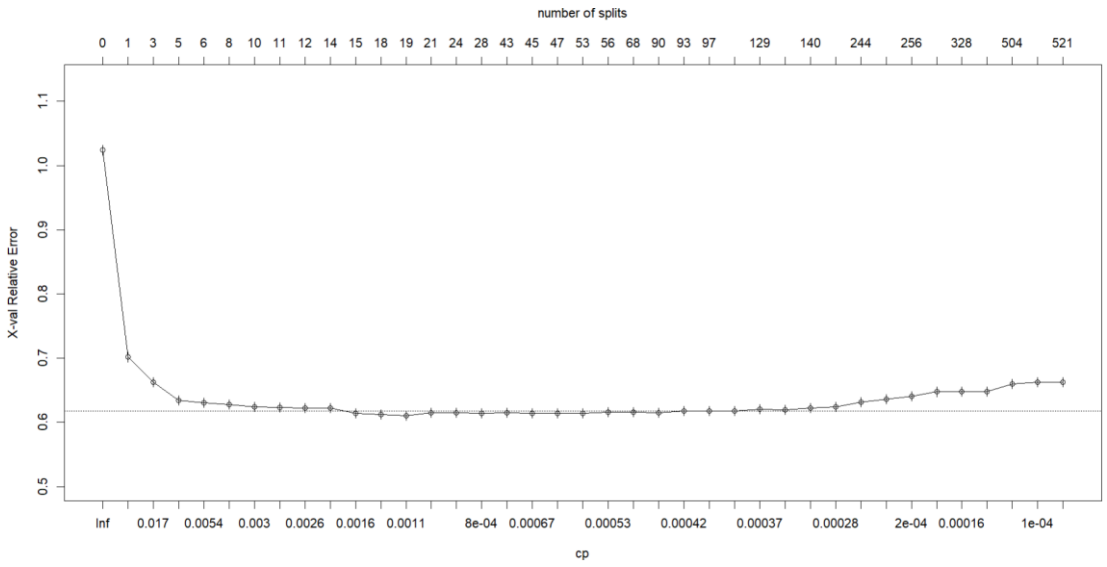
Gràfica 16: Curva ROC on s'hi representa el valor AUC de l'arbre de classificació amb tots els predictors.

En aquest cas, les variables que considera ser més importants són: Gamma GTP, colesterol LDH, edat i sexe. En la gràfica 17 es mostra l'arbre de classificació on en les capses es mostra la categoria predita (Y o N), la proporció d'individus que no pertanyen a la classe predita i la proporció observacions incloses en el node. Observem que comença discriminant pel sexe. En cas de que l'individu sigui masculí la classe predita en aquest node és Y (consumidor), amb una tassa d'error del 68% on la proporció d'observacions incloses en el node es del 54%. En canvi, si el sexe és femení observem que l'error es del 30% sent la classe predita N (no consumidor) i la proporció d'observacions incloses en el node és del 46%.



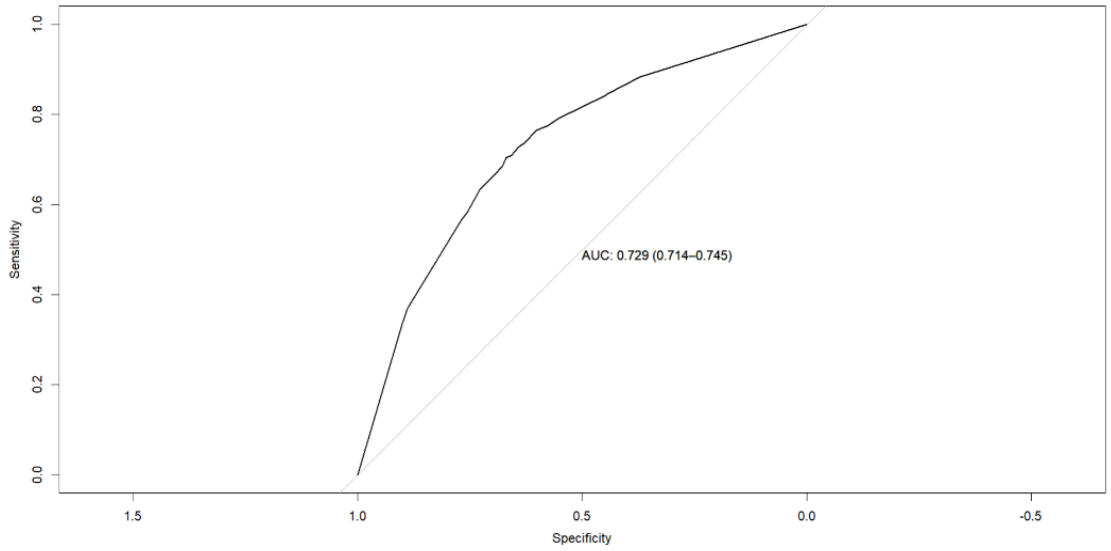
Gràfica 17: Arbre de classificació entrenat amb tots els predictors.

Finalment, tornem a ajustar el model de classificació però utilitzant només els predictors seleccionats pel Random Forest. En aquest cas, el millor model s'assoleix en el valor cp igual a 0.00107950 amb 19 particions i amb un error de 0.61024 (Gràfica 18). Per a aquest model, obtenim un **accuracy de 68.3%**, una sensibilitat de 70.9%, una especificitat de 65.8% i un **AUC** de **72.9%** (Gràfica 19). No es mostra l'arbre, ja que al ser molt gran (tot i tenir només 4 predictors) no permet una visualització decent del contingut.



Gràfica 18: Relació entre l'error i el nombre de particions (i Cp) per

a l'arbre de classificació amb la selecció de predictors.



Gràfica 19: Curva ROC on s'hi representa el valor AUC de l'arbre de classificació

amb els predictors seleccionats pel Random Forest.

Confusion Matrix and Statistics			
		Reference	
		N	Y
Prediction	N	1294	574
	Y	674	1400

Taula 11: Matriu de confusió de l'arbre de classificació

amb la selecció de variables.

Finalment, analitzem la **sensibilitat** i l'**especificitat** dels models. La sensibilitat i l'especificitat són mesures de la capacitat d'una prova per classificar correctament una persona com a consumidora o no. La **sensibilitat** es refereix a la capacitat d'una prova per **designar una persona consumidors com a tal**. Una prova altament sensible significa que hi ha pocs resultats falsos negatius i, per tant, es perden menys casos de consumidors. **L'especificitat** d'una prova és la seva **capacitat per designar un individu que no és consumidors com a no consumidor**. Una prova molt específica significa que hi ha pocs resultats falsos positius. Per tant, lo ideal és tenir valors alts per als dos estadístics.

D'entre els models reduïts, tenim que l'especificitat no varia casi res (Random Forest = 66.7%; Bagged Tree = 65.5%; Arbre de classificació = 65.8%) i en la sensibilitat tenim el valor més alt per a l'arbre de classificació amb un valor de **70.9%** (Taula 12).

Model	Número de predictors	Accuracy	Sensibilitat	Especificitat	AUC
Random Forest	19	71.9%	73.2%	70.6%	80.2%
Random Forest	4	66.2%	65.8%	66.7%	71.9%
Bagged Tree	19	70.6%	71.5%	69.7%	78.1%
Bagged Tree	4	64.7%	63.8%	65.6%	70.5%
Arbre de classificació	19	70.8%	78.3%	63.3%	75.6%
Arbre de classificació	4	68.3%	70.9%	65.8%	72.9%

Taula 12: Taula resum del rendiment dels models.

8. Conclusions

Hem vist que el consum d'alcohol és més freqüent en joves i que a partir del voltant dels 40 anys d'edat, aquest hàbit comença a disminuir i que els homes són els que en consumeixen amb més freqüència en comparació a les dones (homes: 71.5%, dones: 28.5%). També em vist que existeix certa associació entre el consum de tabac i el consum d'alcohol, tenint casi el triple de fumadors (32.6%) i el doble d'ex-fumadors (24.3%) en el grup consumidor respecte al grup no consumidor.

Segons els models ajustats, les variables més essencials són els **nivells de Gamma GTP, nivells de colesterol LDH i nivells d'hemoglobina**. I que a més, ja s'ha vist certa discrepància en les respectives distribucions.

Com que ens interessa tenir un model senzill i la vegada eficient, **l'arbre de classificació** reduït sembla ser el més adequat. Que si el comparem amb el millor model (Random Forest complet) tenim una reducció en la capacitat predictiva, però ho compensa amb el fet de passar de 19 variables a tan sols 4. I un dels avantatges que tenim és el fet que, a diferència de la resta de models, aquest és interpretable.

9. Discussió

Aquests resultats tenen sentit, ja que tal i com hem mencionat a l'inici (Objectius i hipòtesi de l'estudi) els nivells de gamma GTP són més altes en els individus que tenen un consum d'alcohol molt elevat ja que indiquen un mal funcionament del fetge, a més que considera que és la variable més important. Per als nivells d'hemoglobina, també se sap que l'alcohol afecta al metabolisme del ferro, provocant així un augment en els nivells d'hemoglobina o també l'efecte contrari. És a dir, fer decaure els nivells d'hemoglobina. Finalment, també s'ha vist la part positiva del consum, que és augmentar els nivells de colesterol LDH ("colesterol bo"). Per tant, tenim un model coherent.

Pel que fa a la relació entre el sexe/estat fumador i el consum d'alcohol, s'ha de tenir en compte també el context social, ja que a Espanya les conclusions podrien ser totalment diferents.

10. Bibliografia

Doctor Maset Julio (2022). *¿Cómo afecta el alcoholismo al cuerpo?* CinfaSalud. <https://cinfasalud.cinfa.com/p/alcoholismo/> [Consulta: 4/12/2023]

DR. PLA VIDAL JORGE (). *"Alcoholismo. Síntomas, diagnostico y tratamiento. Clínica U. De Navarra (cun.es)" <https://www.cun.es/enfermedades-tratamientos/enfermedades/alcoholismo#:~:text=%C2%BFQu%C3%A9%20es%20el%20alcoholismo%3F,causal%20que%20provoca%*20dicho*%20trastorno%22. >* [Consulta: 4/12/2023].

Servei Nacional d'Assegurança Sanitària de Corea del Sud (2022) *<https://www.data.go.kr/data/15007122/fileData.do>* [Consulta: 15/10/2023]

Sotaquirà Miguel (2021). *"Clasificación con Árboles de Decisión: el algoritmo CART|Codificando Bits".* <https://www.codificandobits.com/blog/clasificacion-arboles-decision-algoritmo-cart/>

González, Juan R. (2022). *"14. Árboles de decisión, Aprendizaje Automático 1."* *<https://isglobal-brge.github.io/Aprendizaje_Automatico_1/index.html >* [Consulta: 20/10/2023]

Consumo de alcohol y perfil lipídcio en participantes del Estudio Longitudinal de Salud del Adulto (ELSA-Brasil) (17 de Febrero del 2020). < https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112019000300024> [Consulta.4/12/2023]

Smith-Garcia, Dorian. (29 de Març del 2023). *"Anemia and Alcohol: Not a Great Mix. Healthline."* <https://www.healthline.com/health/anemia/anemia-and-alcohol> [Consulta:15/12/2023]